X Konferencja z cyklu Diagnostyka edukacyjna
20-22 IX 2004

Roman DOLATA
Wydział Pedagogiczny Uniwersytetu Warszawskiego
Ekspert Instytutu Spraw Publicznych

Materiał przygotowany na X konferencję z cyklu Diagnostyka Edukacyjna: *Teoria i praktyka oceniania zewnętrznego*, 20-22 września 2004, Kraków (fragmenty książki R. Dolaty, E. Putkiewicz, A. Wiłkomirskiej *Reforma egzaminu maturalnego — oceny i rekomendacje*. Warszawa 2004, Instytut Spraw Publicznych)

WYBRANE PROBLEMY RZETELNOŚCI I TRAFNOŚCI NOWEJ MATURY

RZETELNOŚĆ SYSTEMÓW PUNKTACJI ZADAŃ OTWARTYCH

W wypadku arkuszy egzaminacyjnych składających się w całości lub w części z zadań otwartych, warunkiem wstępnym wartościowego oceniania jest rzetelność systemów punktacji, które w postaci liczbowej reprezentują jakość uczniowskiego wytworu. Ponieważ ocenianie zadań otwartych nie jest — w przeciwieństwie do punktacji zadań zamkniętych — procesem zautomatyzowanym, musimy mieć pewność, że procedury punktacji są całkowicie lub przynajmniej w satysfakcjonującym stopniu powtarzalne. Innymi słowy, musimy być pewni, że ocena nie zależy — lub zależy w bardzo niewielkim stopniu — od tego, który z uprawnionych egzaminatorów daną pracę sprawdza. Zadowalający poziom rzetelności punktacji jest warunkiem koniecznym — choć oczywiście niewystarczającym — uznania danego narzędzia egzaminacyjnego za produkt dobrej jakości. Tylko spełnienie tego wstępnego warunku pozwala przystąpić do analiz trafności i rzetelności dydaktycznej oraz psychometrycznej.

W naszym badaniu ewaluacyjnym wykorzystaliśmy prace maturalne pochodzące z nowej matury 2002 i skupiliśmy się na trzech przedmiotach: języku polskim, historii i matematyce. Wybór języka polskiego jest oczywisty, albowiem w nowej formule to jedyny wspólny dla wszystkich maturzystów egzamin, na dodatek narzędzia egzaminacyjne wspierają się na tradycyjnych szkolnych wypracowaniach — z pewnością najbardziej złożonych wytworach szkolnych, które próbujemy oceniać. Liczne badania, również polskie (np. Niemierko 1998, 1999), wskazują, że obiektywne ocenianie wypracowań szkolnych to zadanie bardzo trudne, być może nawet beznadziejne. Tym ważniejsze jest sprawdzenie, w jakim stopnieniu udało się to w 2002 r. Analiza arkuszy z języka polskiego jest priorytetowym zadaniem tego studium.

Egzamin z historii jest interesującym obiektem z dwóch powodów. Przede wszystkim dlatego, że część pierwsza egzaminu (arkusz I) to test składający się głównie z zadań zamkniętych, ale z pewną „domieszką" zadań otwartych. Ciekawe, jak spisują się takie hybrydowe formy. Ponadto część druga (arkusz II) składa się

z odmiennych w stosunku do wypracowań szkolnych, ale też złożonych zadań otwartych — zadań wymagających interpretacji materiałów źródłowych.

Matematyka znalazła się na liście badanych przez nas przedmiotów głównie z powodu pewnej ambiwalencji w zakresie przewidywanego poziomu rzetelności systemów punktacji wykorzystywanych w tym egzaminie. Z jednej strony cele kształcenia matematycznego lepiej poddają się operacjonalizacji, a to dobrze służy rzetelności; z drugiej strony — konstruktorzy arkuszy zdecydowali się na konsekwentne stosowanie zadań otwartych, a to z kolei rodzi problemy z opracowywaniem służących wysokiej rzetelności modeli oceniania. Co z tego wyszło, zobaczymy.

1. METODA

Rzetelność systemów punktacji w oczywisty sposób zależy od poziomu przygotowania egzaminatorów. Jeżeli chcielibyśmy prognozować rzetelność systemów punktacji na maturze 2005 r., powinniśmy pobrać próbkę z całej, kilkudziesięciotysięcznej populacji egzaminatorów. Kłopot w tym, że trudno po blisko dwóch latach od momentu szkoleń dla egzaminatorów uznać, że wszyscy egzaminatorzy są w stanie gotowości. Jak pisaliśmy w poprzednim rozdziale, przed maturą 2005 z pewnością muszą się odbyć szkolenia uzupełniające.

Zdecydowaliśmy się więc zawęzić grupę egzaminatorów, spośród których wylosujemy osoby do naszego badania, tylko do tych, którzy wzięli udział w sprawdzaniu prac w nowej maturze w 2002 r. Z drugiej strony z grupy wyeliminowaliśmy egzaminatorów pracujących w przeszłości lub obecnie w komisjach egzaminacyjnych. W ten sposób nasza próbka egzaminatorów będzie grupą o z pewnością ponadprzeciętnych kompetencjach egzaminacyjnych, a nasze oszacowanie rzetelności będzie raczej oszacowaniem „pułapu", jaki system może osiągnąć przy bardzo dobrym wyszkoleniu egzaminatorów.

Ostatecznie do badania wylosowano po jednym egzaminatorze z każdego okręgu dla każdego z trzech przedmiotów, czyli w sumie 3x8 = 24 egzaminatorów (charakterystykę opinii o nowej maturze wylosowanej próbki egzaminatorów można znaleźć w rozdziale o nauczycielskich opiniach o nowej maturze).

Kluczowym dla wiarygodnego oszacowania rzetelności systemów punktacji zagadnieniem jest reprezentatywność próbki sprawdzanych prac. Kluczowymi parametrami są przeciętny poziom prac i zróżnicowanie tego poziomu. Jeżeli próbki prac znacząco odbiegałyby od całej puli prac, to mogłoby to zafałszować oszacowanie rzetelności. Szczególnie ważne jest odpowiednie zróżnicowanie jakości prac w próbce. Gdy zróżnicowanie prac w próbce jest zdecydowanie niższe niż w całej puli prac — grozi nam niedoszacowanie rzetelności. Gdy zróżnicowanie jest zdecydowanie wyższe — grozi nam przeszacowanie poziomu rzetelności.

Z puli prac archiwalnych nowej matury 2002 roku wylosowano prace z:

- języka polskiego, poziom podstawowy (arkusz I), temat 1;
- języka polskiego, poziom podstawowy (arkusz II);
- języka polskiego, poziom rozszerzony (arkusz III), temat 2;
- historii, arkusz I;
- historii, arkusz II, opcja A;

- matematyki, poziom podstawowy (arkusz I);
- matematyki, poziom rozszerzony (arkusz II).

Arkusze losowano tak, by w próbce były proporcjonalnie reprezentowane prace zdających w 2002 r. nową maturę w poszczególnych okręgach. Docelowo w każdej próbce miało się znaleźć po 50 prac, czyli łącznie 7x50 = 350 prac. By mieć margines swobody, pozwalający wyeliminować z próbek prace nieczytelne lub trudne do skopiowania, losowaliśmy prace z 25% naddatkiem. Okazał się on potrzebny też i z tego powodu, że w jednej komisji okręgowej już nie przechowywano prac z 2002 r., a w drugiej tylko nieliczną, nielosową próbkę. Charakterystyki próbek przedstawia poniższa tabela. Charakterystyki próbek przedstawia poniższa tabela.

Tabela 1. Charakterystyka próbek prac maturalnych

Arkusz	Liczba maturzystów w 2002 r.	Parametry rozkładu ocen prac w 2002 r. średnia/sd	Liczba prac w próbce	Parametry rozkładu ocen prac w próbce średnia/sd
Język polski, poziom podstawowy, arkusz I *	6471	29,0 pkt 11,8	49	30,5 pkt 10,7
Język polski, poziom podstawowy, arkusz II *	6471	18,9 pkt 4,4	50	19,4 pkt 4,9
Język polski, poziom rozszerzony, arkusz III	2155	43,5 pkt 17,5	49	41,8 pkt 15,9
Historia, arkusz I, opcja A **	514	11,6 pkt 5,5	50	13,0 pkt 5,0
Historia, arkusz II, opcja A **	514	29,5 pkt 5,4	50	29,1 pkt 5,4
Matematyka, poziom podstawowy, arkusz I	6468	25,7 pkt 9,8	50	26,5 pkt 10,2
Matematyka, poziom rozszerzony, arkusz II	2208	26,4 pkt 15,8	50	30,1 pkt 15,3

* Parametry dla 2002 r. na podstawie danych uzyskanych z CKE.
** Parametry dla 2002 r. Na podstawie danych uzyskanych z cke, bez prac ocenionych na 0 pkt.

W wypadku doboru prac z historii wyeliminowano z losowania prace ocenione na 0 pkt. Postąpiono tak ponieważ, były to głównie prace „porzucone". Włączenie ich do próbek zawyżyłoby wskaźniki rzetelności. Natomiast z próbek prac z języka polskiego, arkusz I i III, już po badaniu wyeliminowano po jednej pracy, ponieważ okazały się wadliwie dobrane.

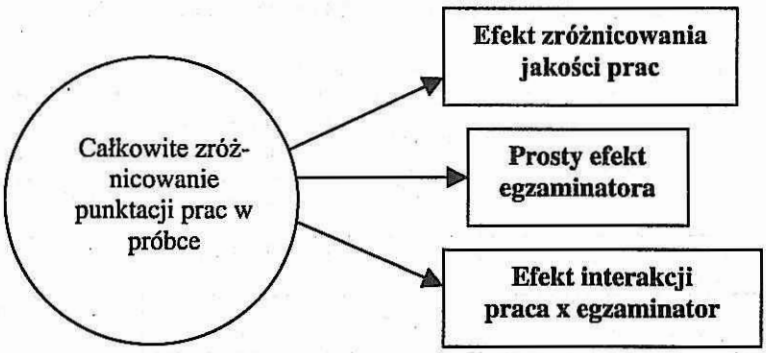
Analiza parametrów rozkładu pokazuje, że można uznać, że próbki są w dostatecznym stopniu reprezentatywne dla puli prac z nowej matury 2002 i mogą być dobrym materiałem do oceny rzetelności systemów punktacji.

Z powodów metodologicznych w naszym badaniu ewaluacyjnym wymagaliśmy pełnej samodzielności. Niektórzy egzaminatorzy zwrócili uwagę, że instrukcja ta nie w pełni przystaje do realiów egzaminacyjnych, ponieważ w 2002 r. mieli oni możliwość konsultowania ocen zarówno z innymi egzaminatorami (procedura nieformalna),

jak i z kierownikiem swojego zespołu egzaminatorów (procedura sformalizowana). Jednak warto zauważyć, że w akcji maturalnej 2005 r. możliwości nieformalnych konsultacji będą ograniczone.

2. MODEL ANALIZY RZETELNOŚCI SYSTEMÓW PUNKTACJI

Poniżej znajduje się model analizy rzetelności systemów punktacji.



Wykres 1. Model analizy rzetelności systemów punktacji: rozkładanie całkowitego zróżnicowania (wariancji) punktacji prac w próbce na efekty składowe

Ponieważ każda praca danego typu (np. prace z historii, arkusz I) była sprawdzana niezależnie przez ośmiu egzaminatorów, mamy macierz 50x8 wyników, czyli łącznie 400 ocen. Punktem wyjścia naszej analizy jest zróżnicowanie tych ocen. W doskonałym systemie punktacji całość tego zróżnicowania (w statystyce używamy pojęcia **wariancji**) wyjaśnia podział na prace, innymi słowy — wszystkie oceny danej pracy są identyczne. Ten właśnie efekt będziemy nazywać efektem zróżnicowania jakości prac. W systemie punktacji o doskonałej rzetelności efekt jakości prac wyjaśnia 100% zróżnicowania ocen. Jeżeli notujemy odstępstwa od tej sytuacji idealnej, mogą być one wynikiem działania jednego z efektów egzaminatora: prostego lub interakcyjnego.

Prosty efekt egzaminatora polega na tym, że każdy z egzaminatorów konsekwentnie — czyli tak samo dla każdej pracy, ale inaczej niż pozostali egzaminatorzy — ustala punkt środkowy skali oceniania. Inaczej mówiąc, jeden egzaminator jest bardziej surowy, inny bardziej szczodry w przydzielaniu punktów. Prosty efekt egzaminatora obniża oczywiście rzetelność, ale jest mniej „groźny" dlatego, że metody jego redukcji są stosunkowo nieskomplikowane.

Trudniejszy do zdefiniowania jest efekt interakcji egzaminator x praca. Można go zdefiniować negatywnie: jeżeli wykluczymy zróżnicowanie ocen pochodne wobec jakości prac oraz wykluczymy zróżnicowane wynikające z prostego efektu egzaminatora, to pozostała wariancja ocen stanowi efekt interakcji. Definiując natomiast ten efekt pozytywnie, możemy powiedzieć, że jest on wynikiem różnego stosowania skal punktowania w odniesieniu do różnych prac. Innymi słowy, kryteria oceny zmieniają się od

pracy do pracy. Efekt interakcji egzaminator x praca jest znacznie bardziej kłopotliwy niż prosty efekt egzaminatora. Po pierwsze, dlatego, że pokazuje na „głębszą" wadliwość sytemu oceniania. Po drugie, sposoby redukcji efektu interakcji są bardziej kosztochłonne.

Do oszacowania efektów jakości pracy i prostego efektu egzaminatora używano jednoczynnikowych analiz wariancji. W pierwszym wypadku zmienną niezależną był numer pracy, w drugim — numer egzaminatora. Miarą siły efektu były wartości wskaźników eta^2 wyrażone w procentach. Efekt interakcji egzaminator x praca szacowano przez dopełnienie do 100% (tzn. jeżeli przykładowo efekt jakości prac wniósł 55%, prosty efekt egzaminatora — 16%, to efekt interakcji egzaminator x praca szacowano na 29%, czyli łącznie 100%). Próbne zastosowanie bardziej złożonych metod szacowania siły efektów — np. model analizy wariancji z powtarzanym pomiarem — przynosiło analogiczne wyniki, zdecydowano się więc na najprostszy model ANOVA.

3. METODY REDUKCJI EFEKTÓW EGZAMINATORA

O rzetelność systemów punktacji należy oczywiście zadbać przede wszystkim w fazie konstrukcji skal. O niektórych zasadach i metodach zapewniania rzetelności systemów punktacji zadań otwartych piszemy w rozdziale o procedurach tworzenia narzędzi egzaminacyjnych. W tym miejscu chcielibyśmy wspomnieć tylko o jednej, choć jak sądzimy, kluczowej metodzie. Rzetelność punktacji powinna być sprawdzana w fazie badań standaryzacyjnych. Stosując metodykę opisaną w tym raporcie należy sprawdzać rzetelność punktacji i gdy jest niesatysfakcjonująca, należy modyfikować system, a gdy to nie przynosi efektu — dany arkusz trzeba usunąć i pracować nad następnym. Można jednak rzetelność systemów punktacji podnieść stosując metody *post factum*, czyli korygując oceny już po przeprowadzeniu egzaminu. Opiszemy i wypróbujemy owocność dwóch takich metod.

Pierwsza metoda zmierza do redukcji prostego efektu egzaminatora. W sytuacji idealnej efekt ten moglibyśmy usunąć, gdyby wszystkie prace były sprawdzane przez jednego egzaminatora i według jednego sytemu punktowania. Oczywiście, w warunkach realnych nie możemy nawet zbliżyć się do tej idealnej sytuacji. Wynika z niej jednak ważna heurystyka zapewniania rzetelności: należy minimalizować liczbę sprawdzających i liczbę stosowanych systemów punktacji. Jeżeli stoimy przed wyborem, czy dać stu egzaminatorom po 100 prac do sprawdzenia, czy pięćdziesięciu po 200 prac, wybieramy drugą możliwość. Oznacza to również, że egzaminatorzy powinni się „wąsko specjalizować" i sprawdzać w danym roku tylko prace danego typu, np. polonista — tylko arkusz I, temat 1. Oznacza to również, że jeżeli nie przemawiają za tym inne ważne względy, należy rezygnować z opcji, np. wyboru tematu wypracowania, okresu historycznego itp.

Pierwsza metoda wykorzystuje efekt tzw. randomizacji. Podobna metoda jest stosowana w Międzynarodowej Maturze. W warunkach realnego egzaminu metoda ta polega na tym, że:

1. Prace egzaminacyjne są losowo przydzielane egzaminatorom. Ważne jest, by pula prac danego typu przedzielona egzaminatorowi była jak największa. Ze względu na efekt randomizacji porcja dla egzaminatora powinna być nie mniejsza niż 100 prac (im więcej, tym lepiej). Najlepiej by było, gdyby prace były dystrybuowane losowo

w obrębie całego kraju. Względy logistyczne wymuszają jednak kompromis — prace należy przydzielać losowo w obrębie okręgu egzaminacyjnego. Najpierw prace są rejestrowane w danej komisji okręgowej, następnie program komputerowo dzieli je losowo na „porcje" dostarczane egzaminatorom.

2. Sprawdzone prace wracają do komisji. Obliczamy średnią ocen wszystkich prac danego typu w okręgu. Następnie obliczamy średnie dla każdego egzaminatora. Zasada randomizacji pozwala nam przyjąć założenie, że średnia jakość prac w każdej porcji była taka sama. Wprowadzamy więc do ocen każdego egzaminatora wskaźniki korekcyjne (*in plus* lub *in minus)* tak, by średnie dla wszystkich porcji były identyczne.

Ten dwuetapowy system można komplikować. Jeżeli niezręczna byłaby formuła ujemnych punktów korekcyjnych, można zastosować taki algorytm korekcji, że dla wszystkich egzaminatorów wskaźniki korekcyjne będą nieujemne. Można również wprowadzić procedury ponownego sprawdzania — przez innych egzaminatorów — tych porcji prac, w których odchylenie od średniej regionalnej jest szczególnie duże. Zasada randomizacji działa, ale pamiętajmy, że ograniczeniem jej skuteczności jest niewielka liczba prac w porcji.

Ważną zaletą tej metody jest jej niewielki koszt.

Druga metoda zmierza do redukcji efektu interakcji egzaminator x praca. W idealnym wydaniu polega ona na tym, że wszystkie prace są sprawdzane przez wszystkich egzaminatorów, a następnie oceny dla każdej pracy są uśredniane. Jakie przybliżenie do tej sytuacji jest realne? Wydaje się, że można jedynie rozpatrywać zdublowanie liczby egzaminatorów. Sama procedura jest prosta. Każda praca jest sprawdzana **niezależnie** przez dwóch egzaminatorów. Egzaminatorzy powinni być losowo dobierani w pary. Jedyny kłopot logistyczny to opracowanie protokołu oceny, który wykluczy zaznaczanie czegokolwiek na pracy, oraz konieczność przekazania porcji prac do drugiego egzaminatora. Egzaminatorzy nie powinni się komunikować. Niezależne oceny są uśredniane i średnia jest ostateczną oceną. Druga metoda częściowo redukuje też prosty efekt egzaminatora, choć przy dwóch egzaminatorach redukcja nie jest duża.

Metodę tę możemy też wykorzystać do racjonalnego rozwiązania problemu „cudu nad kryterium". Rozkłady ocen w nowej maturze 2002 r. wyraźnie pokazują, że bezpośrednio ponad kryterium zaliczenia danego egzaminu maturalnego obserwujemy „wypiętrzenie" rozkładu. Czasami wartości *kryterium plus 1, 2 pkt* to wartości modalne rozkładu. Choć w świetle naszej szkolnej pedagogii oceniania jest to zrozumiałe (ocenianie na korzyść ucznia), to zjawisko to jest nie do pogodzenia z ideą obiektywnego oceniania. Można przyjąć prostą zasadę (oczywiście w obrębie tej metody — czyli sprawdzania zdublowanego), że gdy praca znajduje się nieznacznie „pod" kryterium (np. w przedziale od *kryterium* do *kryterium minus średni rozstęp między ocenami w parach egzaminatorów*), to otrzymuje ocenę nie średnią, ale równą kryterium. Oczywiście należałoby tę metodę połączyć z zakazem manipulowania wynikiem oceny w pobliżu kryterium.

Można by też rozpatrzyć inny wariant metody podwójnego sprawdzania prac, polegający na wystawianiu pracy wyższej z ocen wystawionych przez dwóch egzaminatorów. Ten wariant mniej skutecznie redukowałby efekt interakcji egzaminator x praca, ale miałby „ładną" pedagogię i częściowo rozwiązywał problem „cudu nad kryterium".

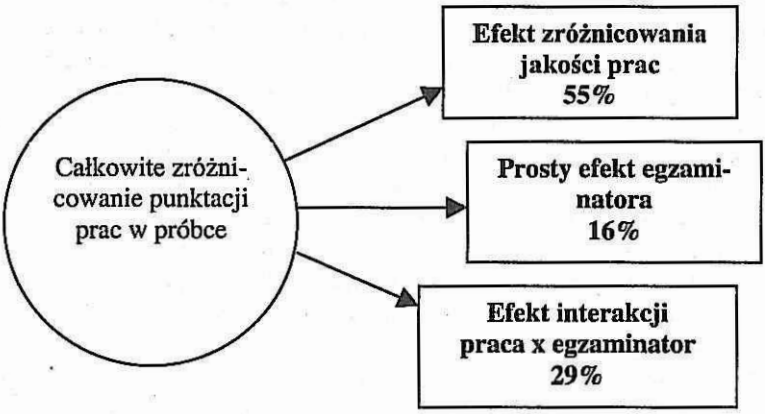
Na zakończenie opisu metody zdublowanego sprawdzania prac rozważmy często postulowaną procedurę ustalania ostatecznej oceny pracy za pomocą negocjowania jej przez sprawdzających egzaminatorów. Na pierwszy rzut oka rozwiązanie to wydaje się najlepsze, zgodne z duchem poszukiwania obiektywnej wartości pracy. Przyjrzyjmy mu się bliżej. Jakie są możliwe wyniki negocjowania ocen w wypadku rozbieżności? Na początku przyjmijmy założenie, że ocena wynikowa będzie w zdecydowanej liczbie przypadków lokować się między pierwotnymi ocenami egzaminatorów. Jeżeli tak, to możliwe są trzy sytuacje: (1) może zwyciężyć ocena jednego z egzaminatorów, (2) mogą spotkać się w pół drogi, (3) dla dobra maturzysty zdecydują się na ocenę korzystniejszą. Zaważmy, że wyniki negocjacji nr 2 i 3 dają te same efekty, co mechaniczne uśrednienie lub przyjęcie zasady wyboru korzystniejszej dla ucznia oceny. Pozostaje efekt nr 1. Jeżeli w każdym przypadku negocjacje zakończą się wynikiem nr 1, to zdublowanie liczby egzaminatorów nie przyniesie redukcji efektu interakcji egzaminator x praca. Drugi problem to zasady negocjowania. Przewidywanie, że negocjacje będą przebiegać w atmosferze zimnej kalkulacji argumentów, jest naiwnością. To raczej będzie złożona społeczna gra o prestiż i „posiadania racji". Przyjęcie, że bardziej wiarygodną oceną pracy jest ocena „dominanta", nie wydaje się racjonalne. Zauważmy wreszcie, że negocjacje wymagają fizycznego kontaktu egzaminatorów, a to może znacząco podwyższać koszty egzaminowania. Podsumowując, strategia negocjowania ostatecznej oceny przez egzaminatorów podwyższa koszty, a owoce są bardzo problematyczne.

Oczywistą wadą zdublowanego punktowania prac maturalnych są koszty. Podwojenie liczby egzaminatorach co prawda ich nie podwaja, ale znacznie zwiększa. Tym ważniejsze jest oszacowanie korzyści ze stosowania tej metody. Taka właśnie symulacja jest jednym z ważniejszych zadań tego badania.

4. ANALIZA RZETELNOŚCI SYSTEMU PUNKTACJI ARKUSZA I Z JĘZYKA POLSKIEGO

Zgodnie z planem analizy zaczynamy od wypracowania na poziomie podstawowym, czyli od arkusza I z języka polskiego. Do analizy wybrano prace na temat 1. Wybrała go większość zdających. Temat brzmiał: *Jaką funkcję pełni Mazurek Dąbrowskiego w „Panu Tadeuszu" Adama Mickiewicza? W pracy wykorzystaj znajomość utworu Józefa Wybickiego i podanych fragmentów „Pana Tadeusza".*

Wypracowania sprawdzano przy użyciu pięciu skal. Pierwsza — kryterialna — dotyczyła rozwinięcia tematu, cztery pozostałe — szacunkowe — służyły do oceny kolejno: kompozycji, stylu, poprawności językowej i szczególnych walorów pracy. [Szczegółowe analizy punktowania prac według tych kryteriów może Czytelnik znaleźć w książce, na której oparto referat.] Efekty analizy rzetelności skali sumarycznej dla arkusza I z języka polskiego przedstawia poniższy wykres.



Wykres 2. Oszacowanie efektu zróżnicowania jakości prac, efektu egzaminatora oraz efektu interakcji praca x egzaminator dla skali sumarycznej, język polski, arkusz I

Gdyby przewidywać rzetelność skali sumarycznej na podstawie rzetelności skal składowych, należałoby szacować ją na 35–40%. Okazuje się jednak, że wskaźnik sumaryczny ma wyższą rzetelność — 55%. Jak to wyjaśnić? Przede wszystkim wynika to z faktu, że proste efekty egzaminatora nie sumują się, a raczej znoszą. Innymi słowy, egzaminator surowo oceniający na jednej skali jest bardziej hojny w stosowaniu innej skali. Oczywiście, to znoszenie się nie jest doskonałe i w sumie prosty efekt egzaminatora wynosi 16%, ale to znacząco mniej, niż wynosi średnia tego efektu dla poszczególnych skal. Po drugie, wyższa rzetelność skali sumarycznej może wynikać — podobnie jak w wypadku skali rozwinięcia tematu — z odwrócenia „logiki" oceniania, czyli uzgadniania sumarycznej liczby punktów z intuicyjną, globalną oceną pracy.

Oczywiście 55-procentowa rzetelność nie może w żadnym stopniu zadowalać. Oznacza ona, że ocena pracy w 45% zależy od czynników subiektywnych. Jak to się przykłada na ocenę poszczególnych prac? Efekt egzaminatora, mierzony odchyleniem ocen dokonywanych przez ośmiu egzaminatorów dla poszczególnych prac, waha się od 4,2 do 12,1 pkt. Średnia odchyleń standardowych wynosi 7,8 pkt. W miarach rozstępu (maksymalna minus minimalna ocena danej pracy) efekt ten przybiera wartość od 11 do 38 pkt, ze średnią 22,3 pkt. Średni rozstęp stanowi 32% teoretycznej zmienności wskaźnika sumarycznego (0–70 pkt). Widzimy więc, że 55-procentowa rzetelność skali sumarycznej przekłada się na rażąco duże rozbieżności w ocenach prac przez poszczególnych egzaminatorów.

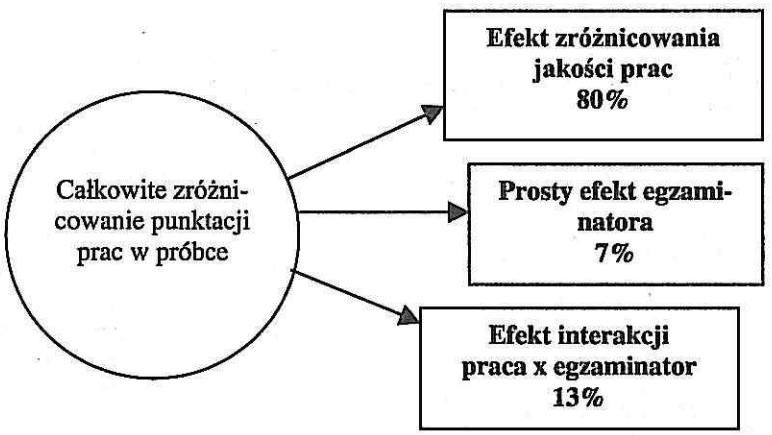
5. ANALIZA RZETELNOŚCI SYSTEMU PUNKTACJI ARKUSZA II Z JĘZYKA POLSKIEGO

Na poziomie podstawowym maturzyści oprócz pisania wypracowania rozwiązywali test sprawdzający rozumienie czytanego tekstu. Jak pisaliśmy w rozdziale poświęconym procedurom budowy narzędzi egzaminacyjnych do matury 2002 r., główną wadą tego testu jest skrajna nierzetelność dydaktyczna. Przypomnijmy, błąd próbkowania celu kształcenia, jakim jest rozumienie tekstów, jest w tym wypadku maksymalny.

W arkuszu II egzaminu z języka polskiego wykorzystano fragment tekstu Romana Ingardena *Książeczka o człowieku.* Rozumienie badano za pomocą 15 zadań krótkiej odpowiedzi.

Analizę rzetelności sytemu punktacji rozpoczniemy od oszacowania efektu zróżnicowania jakości prac, efektu egzaminatora oraz efektu interakcji praca x egzaminator dla poszczególnych zadań. Poszczególne zadania były oceniane na skalach szacunkowo-kryterialnych. Modele odpowiedzi przewidywały jako ocenę maksymalną od 1 do 4 pkt. [Szczegółowe analizy — jak poprzednio.]

Wyniki analizy rzetelności przedstawia wykres 3.



Wykres 3. Oszacowanie efektu zróżnicowania jakości prac, efektu egzaminatora oraz efektu interakcji praca x egzaminator wskaźnik rozumienia tekstu, język polski, arkusz II

Rzetelność punktacji testu rozumienia tekstu jest wyższa niż rzetelność punktacji wypracowań. Osiemdziesięcioprocentowy wskaźnik rzetelności jest dobrym prognostykiem dla możliwości budowania dobrych testów rozumienia tekstów wykorzystujących obficie zadania krótkiej odpowiedzi. Oczywiście nie oznacza to, że test z roku 2002 jest już takim narzędziem. Wystarczy przejrzeć wyniki szczegółowe, by zrozumieć, że nie sposób zadowolić się 80-procentową rzetelnością. Efekt egzaminatora mierzony wielkością odchylenia standardowego ocen wystawionych poszczególnym pracom waha się od 1,1 do 3,7 pkt. Średnia odchyleń standardowych wynosi 2,0 pkt. Efekt egzaminatora wyrażony wielkością rozstępu (maksymalna minus minimalna

liczba punktów przyznana danej pracy przez egzaminatorów) przebiera wartość od 3 do 11 pkt, ze średnią 5,7 pkt. Średni rozstęp stanowi 19% teoretycznej zmienności wskaźnika sumarycznego (0–30 pkt).

Maksymalną wielkość efektu egzaminatora notujemy dla pracy 21 — odchylenie standardowe ocen wynosi 3,7 pkt. Oznacza to przy pierwszym sposobie kalkulowania 35% maksymalnej wartości, przy drugim — 51% maksymalnej wartości odchylenia standardowego. W wypadku najmniej „kontrowersyjnej" pracy procenty wynoszą odpowiednio 10 i 15 punktów procentowych wartości maksymalnej. To oczywiście zdecydowanie niższe nasilenie efektów egzaminatora niż w wypadku wypracowań. Ale przynajmniej w wypadku prac o relatywnie wysokim nasileniu efektu egzaminatora jest to w dalszym ciągu poziom zbyt duży.

6. PODSUMOWANIE ANALIZY RZETELNOŚCI PUNKTOWANIA PRAC Z JĘZYKA POLSKIEGO

Spróbujmy podsumować wyniki analizy rzetelności systemów punktacji dla arkuszy egzaminacyjnych z języka polskiego. Komplet wyników przedstawia poniższa tabela.

Tabela 2. Zestawienie oszacowań rzetelności punktowania dla poszczególnych skal i arkuszy z języka polskiego. Procent wariancji wyjaśnianej przez dany czynnik

Skala lub arkusz	Efekt zróżnicowania jakości prac	Prosty efekt egzaminatora	Oszacowany efekt interakcji praca x egzaminator
Arkusz I łącznie (wypracowanie, poziom podstawowy)	55	16	29
Skala rozwinięcia tematu	47	21	32
Skala kompozycji	36	24	40
Skala stylu	35	11	54
Skala poprawności językowej	44	18	38
Skala szczególnych walorów pracy	22	21	57
Arkusz III łącznie (wypracowanie, poziom rozszerzony)	49	22	29
Skala rozwinięcia tematu	54	19	27
Skala kompozycji	33	18	49
Skala stylu	31	18	51
Skala poprawności językowej	34	21	45
Skala szczególnych walorów pracy	23	24	53
Arkusz II (rozumienie tekstu)	80	7	13

Jak widzimy, systemy punktacji wypracowań — zarówno dla poziomu podstawowego, jak i rozszerzonego — mają bardzo niską, zdecydowanie nie satysfakcjonującą rzetelność. Oszacowania siły poszczególnych efektów pozwalają stwierdzić, że ocena punktowa pracy maturzysty w mniej więcej w tym samym stopniu zależy od jej cech,

co od cech egzaminatora. Stosowanie takich systemów oceniania oznaczałoby, że stronniczość oceniania przez szkolnego nauczyciela zastąpiliśmy nie mniejszym „losowym szumem" związanym z efektami egzaminatora.

Za niską rzetelność systemów punktacji wypracowań odpowiadają przede wszystkim skale szacunkowe. Spośród nich szczególnie katastrofalną rzetelność ma skala szczególnych walorów pracy. Wynik ten nie dziwi. Sam charakter skali czynił z niej usankcjonowaną „oazę" subiektywności, ale nie zmienia to faktu, że użycie takiej skali nie da się pogodzić z dążeniem do obiektywizmu. Również skala jakości kompozycji i stylu w obecnej postaci nie może być używana.

Rozczarowuje też skala rozwinięcia tematu. Jej kryterialny charakter mógł dawać nadzieję na zdecydowanie wyższą rzetelność. Jest ona wyższa niż skal szacunkowych, ale i tak zdecydowanie poniżej poziomu minimum przyzwoitości.

Opisane w poprzednich rozdziałach raportu metody naprawy systemu punktacji wypracowań (standaryzacja w badaniach pilotażowych, inne definiowanie skal szacunkowych) być może pozwolą poprawić rzetelność punktowania. Ale trudno spodziewać się uzyskania w ten sposób w pełni satysfakcjonującego poziomu rzetelności. By poziom ten przekroczył próg „przyzwoitości", konieczne będzie zastosowanie metod *post factum* redukcji efektu egzaminatora. Zobaczymy w kolejnym podrozdziale, jakie mogą być efekty stosowania tych metod.

Na tle systemów punktacji wypracowań, procedury punktacji testu czytania ze zrozumieniem wypadają lepiej, choć nie znaczy to, że dobrze. Efekty egzaminatora to w tym wypadku 20% zmienności ocen, przy czym 7% to prosty efekt egzaminatora, relatywnie łatwy do wyeliminowania. Musimy jednak pamiętać o dwóch sprawach. Po pierwsze, swą wysoką rzetelność skala zawdzięcza głównie tym zadaniom skali, które odnoszą się do najniższych kategorii taksonomicznych. Zadania głębiej wnikające w rozumienie tekstu zdecydowanie trudniej poddawały się rzetelnemu punktowaniu. Po drugie, arkusz II obarczony jest błędem rzetelności dydaktycznej związanym ze skrajnie dużym błędem „próbkowania" (tylko jeden tekst!). Niemniej wydaje się, że zwiększenie liczby tekstów, zastąpienie części pytań krótkiej odpowiedzi zadaniami zamkniętymi oraz dopracowanie modeli odpowiedzi w wypadku zadań odnoszących się do wyższych kategorii taksonomicznych pozwoli zbudować satysfakcjonujący test rozumienia czytanych tekstów. Jeżeli do tego zastosujemy opisane we wstępie do tej części raportu, a zastosowane poniżej, metody redukcji efektów egzaminatora, to być może otrzymamy dobry test ze względu na rzetelność punktowania.

7. SYMULACJA KORZYŚCI Z ZASTOSOWANIA METOD *POST FACTUM* PODWYŻSZANIA RZETELNOŚCI PUNKTACJI ARKUSZY Z JĘZYKA POLSKIEGO

Jak można poprawić jakość systemów punktowania? Oczywiście przede wszystkim przestrzegając zasad dobrej budowy skal i sprawdzając rzetelność w fazie badań standaryzacyjnych. Ale możliwe jest też stosowanie dwóch metod podwyższania rzetelności punktowania zadań otwartych, które możny stosować *post factum*. Pierwsza nastawiona na eliminacje prostego efektu egzaminatora polega na wprowadzeniu współczynników korekcyjnych, druga nastawiona na redukcję efektu interakcji egza-

minator x praca polega na sprawdzaniu każdej pracy przez większą liczbę egzaminatorów.

Pierwsza metoda jest prosta i — przypomnijmy — polega na wprowadzeniu wskaźników korekcyjnych, sprawiających, że znika prosty efekt egzaminatora. W wypadku naszych badań po prostu do każdej oceny możemy dodać lub odjąć wskaźnik „surowości" obliczony dla poszczególnych egzaminatorów. Na przykład, jeżeli w wypadku testu rozumienia czytanego tekstu średnia ocen egzaminatora nr 8 jest o 2,3 pkt niższa od średniej ogólnej, to do wszystkich ocen tego egzaminatora dodajemy 2,3 pkt. W ten sposób średnie ocen dla wszystkich egzaminatorów są równe. W warunkach prawdziwego egzaminu — przypomnijmy — strategia ta bazuje na efekcie randomizacji i wymaga, by egzaminatorzy sprawdzali dość liczne, losowe próbki prac.

Efekty zastosowania tej metody przedstawia tabela 3. Oczywiście, prosty efekt egzaminatora równa się w tym wypadku zero.

Tabela 3. Zestawienie oszacowań rzetelności punktowania dla poszczególnych arkuszy z języka polskiego po wytrąceniu prostego efektu egzaminatora. Procent wariancji wyjaśnianej przez dany czynnik

Skala lub arkusz	Efekt zróżnicowania jakości prac	Oszacowany efekt interakcji praca x egzaminator
Arkusz I łącznie (wypracowanie, poziom podstawowy)	65	35
Arkusz III łącznie (wypracowanie, poziom rozszerzony)	63	37
Arkusz II (rozumienie tekstu)	87	13

Efekty zastosowania korekcji są zróżnicowane. Największą poprawę obserwujemy w wypadku arkusza III. Zysk w tym wypadku wynosi 14 punktów procentowych. W wypadku arkusza I notujemy przyrost wskaźnika rzetelności rzędu 10 punktów, w wypadku arkusza I — 7 punktów procentowych. Widzimy więc, że im silniejszy prosty efekt egzaminatora, tym większego możemy się spodziewać zysku. To wniosek oczywiście banalny, ale w jakimś sensie paradoksalny. Pokazuje on bowiem — o czym pisaliśmy już wcześniej — że prosty efekt egzaminatora jest zdecydowanie mniej „szkodliwy", bo jest w miarę łatwy do usunięcia.

Drugi wniosek wynika z pierwszego. Widzimy, że przy stosunkowo niewielkich nakładach możemy uzyskać znaczącą, choć oczywiście dalece jeszcze niewystarczającą, poprawę rzetelności. Stworzenie odpowiednich procedur korekcyjnych, które mogłyby być wykorzystane w realnych warunkach egzaminacyjnych, jest możliwe i wymaga tylko opracowania zasad dystrybucji prac do sprawdzania oraz — co było już przez nas postulowane również z innych powodów — wprowadzenia zamiast sztywnej skali procentowej, znacznie bardziej elastycznej formuły wystawiania ostatecznej oceny maturalnej w postaci punktów przeliczeniowych.

Czy 63–65% poziom rzetelności punktowania w wypadku oceny wypracowań jest dobrym wynikiem? Oczywiście, w miarach bezwzględnych to w dalszym ciągu poziom katastrofalnie niski. Ale z czym możemy ten poziom porównać, jaki punkt odniesienia przyjąć, by ocenić owoce podjętego trudu obiektywizacji oceny tak złożonych wytworów, jak wypracowania szkolne? Wydaje się, że naturalnym punktem odniesienia byłby poziom rzetelności uzyskiwany przy stosowaniu tradycyjnych metod szkolnego oceniania. Przywoływane we wstępie badania Niemierki (1998, 1999) sugerują, że systemy oceniania kryterialnego nie są wiele lepsze od tradycyjnego oceniania. Nie dysponujemy jednak oszacowaniami rzetelności tradycyjnego oceniania szkolnego, dokonanymi zgodnie z przyjętą przez nas metodologią. Jednak — uprzedzając fakty opisywane w następnym rozdziale — możemy skorzystać z danych z analizy trafności. Jeden zastaw sędziów kompetentnych składał się z czterech nauczycieli. Mieli oni te same prace, które zostały użyte do analizy rzetelności punktacji, ocenić globalnie, na modłę szkolną, porządkując wypracowania od najlepszego do najsłabszego. Efekty porządkowania przedstawiono w postaci rang. Okazało się, że poziom rzetelności (zgodność uporządkowań) oszacowany za pomocą modelu analizy wariancji wyniósł zarówno dla arkusza I, jak i III 62%. To wynik bardzo zbliżony do poziomu uzyskanego powyżej. Dodajmy, że w grupie sędziów kompetentnych — pracowników uniwersytetu — analogiczne wskaźniki były zdecydowane niższe i wyniosły odpowiednio 47 i 48%. Okazuje się więc, że skomplikowane zabiegi zmierzające do obiektywizacji oceniania są prawie bezowocne. „Szkolna kultura oceniania" daje podobną zbieżność ocen, jak rozbudowany system punktacji. Czy nie można rzetelniej? Zobaczmy, co przyniesie podwojenie liczby egzaminatorów.

Przejdźmy do symulacji efektów zastosowania drugiej metody podwyższania rzetelności nastawionej na redukcję efektu interakcji. Jak pamiętamy, efekt interakcji praca x egzaminator jest znacznie trudniej redukowalny. W idealnej sytuacji wszystkie prace powinny być sprawdzane przez jednego egzaminatora lub każda praca powinna być sprawdzana przez wszystkich egzaminatorów, a wynik uśredniany. Z oczywistych względów jedyna realna strategia to niezależne sprawdzanie prac przez dwóch egzaminatorów i uśrednienie oceny. Wypróbujemy więc, jaki zysk w poziomie rzetelności dałoby **niezależne** sprawdzanie każdej pracy przez dwóch, losowo dobranych egzaminatorów. Przed symulacją z ocen wytrącony już został prosty efekt egzaminatora.

Losowo dzielimy ośmiu egzaminatorów na dwie grupy. Oznaczmy je umownie grupą A i B. Następnie każdemu egzaminatorowi z grupy A losowo przydzielamy egzaminatora z grupy B. Oszacujemy rzetelność punktowania, czyli efekt jakości pracy trzykrotnie: dla grupy A, dla grupy B oraz dla ocen będących uśrednieniem ocen w parach. Wylosowane pary to (pierwszy egzaminator — grupa A, drugi w parze — grupa B): egzaminatorzy nr 1 i 7, 2 i 5, 6 i 3 oraz 8 i 4. Zgodność ocen w parach była dość zróżnicowana. Dla arkusza I korelacje w parach wahały się od r=0,63 dla pary egzaminatorów nr 6 i 3 do r=0,71 dla pary egzaminatorów nr 1 i 7; dla arkusza III korelacje w parach wahały się od r=0,42 dla pary egzaminatorów nr 6 i 3 do r=0, 67 dla pary egzaminatorów nr 2 i 5; dla arkusza II korelacje w parach wahały się od r=0,77 dla pary egzaminatorów nr 6 i 3 do r=0,89 dla pary egzaminatorów nr 2 i 5.

Tabela 4. Zestawienie oszacowań rzetelności punktowania dla poszczególnych arkuszy z języka polskiego przez dwóch egzaminatorów. Procent wariancji wyjaśnianej przez czynnik jakości pracy

Skala lub arkusz	Efekt zróżnicowania jakości prac w grupie A	Efekt zróżnicowania jakości prac w grupie B	Efekt zróżnicowania jakości prac dla ocen uśrednionych dla par A-B
Arkusz I łącznie (wypracowanie, poziom podstawowy)	74	62	79
Arkusz III łącznie (wypracowanie, poziom rozszerzony)	67	65	80
Arkusz II (rozumienie tekstu)	91	87	94

Analizę wyników zacznijmy od arkusza I. Ponieważ wyjściowy poziom rzetelności w grupie A i B bardzo się różnią, za punkt wyjścia przyjmijmy średnią oszacowań, czyli 68%. Porównując tę wartość z poziomem rzetelności dla średnich punktacji obliczonych w parach widzimy, że zysk wyniósł 11 punktów procentowych. Analogicznie dla arkusza III zysk ten wyniósł 14 punktów, dla arkusza II — 5 punktów. Widzimy więc, że w wypadku sprawdzania wypracowań zysk ze zdublowania liczby egzaminatorów jest znaczący, a w wypadku testu rozumienia tekstu — mniejszy. Ale też w tym ostatnim wypadku z pewnością trudniej o postęp. Startowanie z poziomu blisko 90% oznacza, że każdy punkt znacząco przybliża nas do pożądanego poziomu.

Podsumowując można powiedzieć, że niezależne sprawdzanie wypracowań przez dwóch egzaminatorów przyniosło trudne do pogardzenia efekty. Oczywiście, ok. 80% rzetelność nie może nas zadowolić, ale jeżeli weźmiemy pod uwagę korzyści, jakie możemy uzyskać rezygnując z jawnie subiektywistycznych skal, dopracowując pozostałe skale oraz rozwijając kulturę oceniania kryterialnego poprzez doskonalenie sytemu szkoleń i procedur kontrolnych w trakcie realnego sprawdzania prac, to możemy z dość dużym prawdopodobieństwem prognozować osiągnięcie 90–95% rzetelności. To jeszcze nie ideał, ale poziom, który może być zaakceptowany. W świetle tych ustaleń wydaje się, że zdublowanie liczby egzaminatorów jest racjonalne i wzrost kosztów ma szanse przełożyć się na znaczący postęp w zakresie obiektywności oceniania wypracowań maturalnych z języka polskiego.

Do trochę innych wniosków musimy dojść w wypadku testu rozumienia czytanych tekstów. Osiągnięty w naszej symulacji poziom rzetelności już spełnia kryterium minimum. Ale pojawia się pytanie, czy w tym wypadku analogicznego efektu nie można uzyskać niższym kosztem? Wydaje się, że odpowiedź jest twierdząca. Zwiększenie liczby tekstów (konieczna ze względu na rzetelność dydaktyczną), wprowadzenie tam, gdzie nie obniża to trafności, zadań zamkniętych, dopracowanie zadań krótkiej odpowiedzi, a szczególnie modeli odpowiedzi do tych zadań (oceny eksperckie modeli, ale i empiryczna analiza w badaniach standaryzacyjnych), obiecuje osiągnięcie podobnego

poziomu rzetelności przy zdecydowanie niższych kosztach. Gdy te tańsze metody zapewniania rzetelności systemów punktacji zastosujemy i przyniosą one spodziewane efekty, być może warto będzie — gdy będzie nas na to stać — rozważyć podwojenie liczy sprawdzających, by wyśrubować poziom rzetelności do pożądanego poziomu 99–100%.

Na zakończenie zauważmy, że w naszej symulacji zastosowanie obu metod *post factum* poprawy rzetelności przyniosło w sumie znaczące podwyższenie rzetelności. Dla systemów punktacji wypracowań zysk wyniósł odpowiednio 24 i 31 punktów procentowych, dla testu rozumienia tekstu — 14 punktów procentowych.

8. ANALIZA RZETELNOŚCI SYSTEMU PUNKTACJI ARKUSZY Z HISTORII — PODSUMOWANIE

Wyniki podsumowujące analizę rzetelności punktacji dla arkuszy z historii przedstawia tabela 5.

Tabela 5. Zestawienie oszacowań rzetelności punktowania dla poszczególnych arkuszy z historii. Procent wariancji wyjaśnianej przez dany czynnik

Skala lub arkusz	Efekt zróżnicowania jakości prac	Prosty efekt egzaminatora	Oszacowany efekt interakcji praca x egzaminator
Arkusz I (test)	95	2	3
Arkusz II (interpretacja tekstów źródłowych)	58	29	13

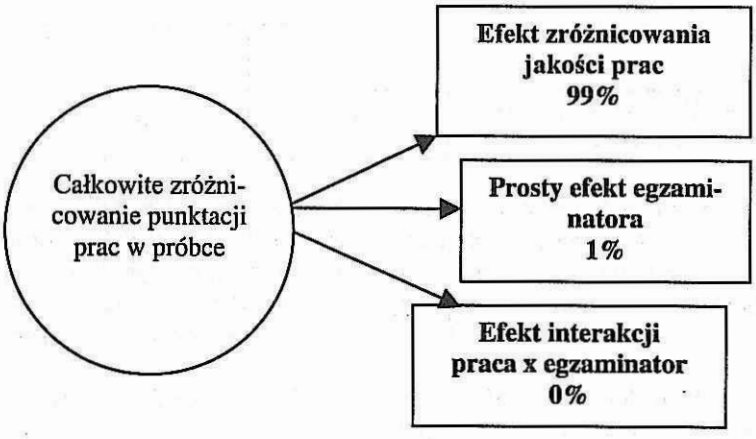
System punktacji testu historycznego ma sumaryczną rzetelność zbliżoną do arkuszy matematycznych (patrz następne podrozdziały). W wypadku testu główną metodą poprawy rzetelności powinno być dopracowanie modeli odpowiedzi dla pytań krótkiej odpowiedzi i/lub ograniczanie liczby tego typu pytań do sytuacji, gdzie wymaga tego trafność. Przypomnijmy, wszędzie tym, gdzie bez szkody dla trafności można stosować pytania zamknięte, należy to robić.

W wypadku arkusza II notujemy wyniki porównywalne do rzetelności systemów punktacji wypracowań z języka polskiego, więc i w tym wypadku należałoby powiedzieć, że stosowanie takich systemów oceniania oznaczałoby, że stronniczość oceniania przez szkolnego nauczyciela zastąpiliśmy „losowym szumem" związanym z efektami egzaminatora.

9. ANALIZA RZETELNOŚCI SYSTEMU PUNKTACJI ARKUSZA I Z MATEMATYKI

Analizę rzetelności systemów punktacji zakończymy zbadaniem skal użytych w obszarze osiągnięć matematycznych. Rokowania w wypadku tych narzędzi egzaminacyjnych są optymistyczne. Wydaje się, że mimo wykorzystania dość złożonych zadań otwartych, „definiowalny" charakter kryterium poprawności rozwiązań pozwala przewidywać wysoką zgodność ocen.

Arkusz I — poziom podstawowy — zawierał 10 zadań ocenianych kryterialnie i punktowanych na skalach od 0 do 3, 4 lub 5 pkt. W sumie można było zdobyć od 0 do 40 pkt. W analizie ograniczymy się do oszacowania efektów dla całych skal, a nie poszczególnych kryteriów/czynności. Nie można poszczególnych kryteriów/czynności potraktować jako niezależnych zmiennych z dwóch powodów. Po pierwsze, czynności wyróżnione w modelach odpowiedzi nie były logicznie niezależne, co oznacza, że zmienne zdefiniowane przez te czynności też są zależne. Innymi słowy, w wielu wypadkach wykonanie lub niewykonanie jednej czynności warunkowało wykonanie innej lub innych czynności w obrębie modelu odpowiedzi. To oczywiście nie jest wadą, ale uniemożliwia traktowanie czynności jako niezależnych — w sensie statystycznym — zmiennych. Nie można poszczególnych kryteriów/czynności potraktować jako niezależnych zmiennych również dlatego, że egzaminatorzy mogli stosować przy ocenie własne, alternatywne modele rozwiązania zadania. Jedynym warunkiem utworzenia alternatywnego modelu dla rozwiązania danego zadania była ekwiwalentność punktowa.



Wykres 4. Oszacowanie efektu zróżnicowania jakości prac, efektu egzaminatora oraz efektu interakcji praca x egzaminator, wskaźnik osiągnięć matematycznych, poziom podstawowy — arkusz I

Widzimy więc, że arkusz I tworzy skalę niemal doskonałą pod względem rzetelności punktowania.

Ponieważ rzetelność sytemu punktacji dla arkusza I jest wysoka, nie będziemy analizować efektu egzaminatora dla wszystkich prac. Należy jednak zdać sobie sprawę, że 99-procentowa rzetelność sytemu punktacji nie oznacza całkowitego braku rozbieżności w ocenie prac. Rozbieżności są oczywiście nieporównanie mniejsze niż w wypadku systemów punktacji dla języka polskiego czy dla arkusza II z historii, ale i tak w wypadku kilku prac dość duże. Na przykład dla pracy rekordzistki nr 31 rozstęp między oceną maksymalną a minimalną wynosi 7 pkt, dla 6 prac rozstęp wynosi 5–6 pkt, dla 22 prac wynosi 3–4 pkt, dla 19 prac wynosi 1–2 pkt i w końcu dla 2 prac notujemy całkowitą zgodność ocen. Średni rozstęp wynosi 2,7 pkt, co stanowi 7% teoretycznej zmienności wskaźnika sumarycznego (0–40 pkt). Również analiza odchyleń standardowych ocen pokazuje, że maksymalną wielkość efektu egzaminatora obserwujemy w wypadku pracy 31 — odchylenie standardowe ocen wynosi 2,2 pkt, przy średniej punktacji 20,9 pkt. Oznacza to przy pierwszym sposobie kalkulowania 10% maksymalnej wartości odchylenia standardowego, przy drugim — 16% maksymalnej wartości rozbieżności ocen.

10. ANALIZA RZETELNOŚCI SYSTEMU PUNKTACJI ARKUSZA II Z MATEMATYKI

Arkusz II — poziom rozszerzony — zawierał 9 zadań ocenianych kryterialnie i punktowanych na skalach od 0 do 4, 5, 6, 7, 8 lub 10 pkt. Numeracja zadań jest wspólna z numeracją arkusza I, w związku z tym pierwsze zadanie arkusza II nosi numer 11, itd. W arkuszu II w sumie można było zdobyć od 0 do 60 pkt. W analizie — analogicznie jak wypadku arkusza I — ograniczymy się do oszacowania rzetelności punktowania dla całych zadań, a nie dla poszczególnych kryteriów/czynności.

W wypadku poziomu rozszerzonego rzetelność punktacji dla poszczególnych zadań wygląda trochę gorzej niż dla arkusza I. Szczególnie niepokoić musi zadanie 12. Zdecydowanie warto dokonać wnikliwej analizy zarówno samego zadania, jak i modelu odpowiedzi, by zrozumieć przyczyny nierzetelności. Również trochę niepokoić musi niższa rzetelność zadania 11 i 16. Wątpliwości co do rzetelności punktowania w arkuszu II mieli niektórzy egzaminatorzy biorący udział w badaniu. Jeden z nich w uwagach napisał:

Porównując kilka próbnych zestawów z maturą 2002 z matematyki stwierdzam, że schemat punktowania dla arkusza II był w tym zestawie wyjątkowo nieprecyzyjny i stwarzał nieznośną dowolność w przyznawaniu punktów. Np. w zadaniu 17 — przypuśćmy, że uczeń nieprawidłowo rozwiązał równanie trygonometryczne, następnie źle wyznaczył moce zbiorów A i Ω, po czym poprawnie obliczył P(A)= A/Ω. Przyznać punkt czy nie? Umowa była: nie przyznawać, ale nie mam pewności, czy w całym kraju oceniano tak samo.

Jak widzimy, obawy w wypadku zadania 17 nie potwierdziły się, ale ogólny niepokój co do jednoznaczności modeli odpowiedzi do arkusza II częściowo znalazł potwierdzenie. To tylko upewnia nas, że konieczne jest sprawdzanie modeli odpowiedzi w trakcie badań standaryzacyjnych na znacznym próbkach prac.

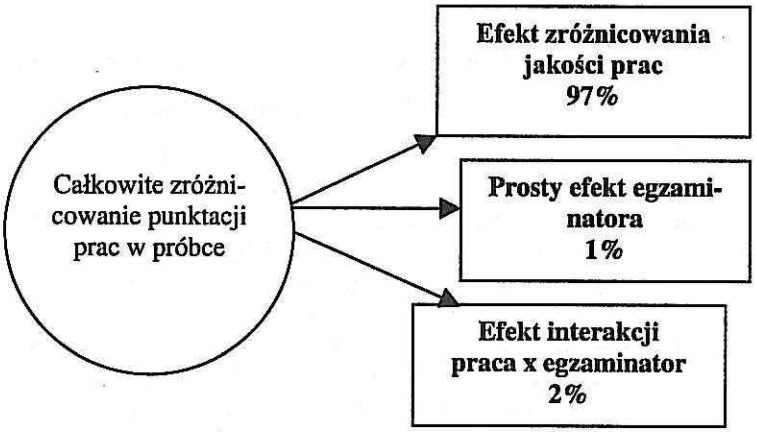
Pewne światło na przyczyny fatalnej — w świetle dobrych wyników dla innych zadań — rzetelności zadania 12 rzucają uwagi innego egzaminatora. W uwagach do treści zadań pisze:

[Uwaga] *dotyczy składu — choć oznaczenie zdarzenia warunkowego A/B jest poprawne, to jednak wprowadzany symbol zdarzenia warunkowego wygląda inaczej — jest kreską pionową prostą a nie pochyłą. Zdarzało się, że absolwent traktował oznaczenie A/B jako różnice zbiorów.*

Łącznie jednak zadania arkusza II mają szansę stworzyć dość rzetelną, ze względu na system punktacji, skalę.

Przyjrzyjmy się parametrom rozkładu. Parametry rozkładu wynoszą: średnia 28,1, odchylenie standardowe 14,1, skośność 0,02. Choć rozkład odbiega od normalnego, to kluczowy dla analizy wariancji parametr — skośność — pozwala z zaufaniem sięgnąć po model analizy wariancji.

A oto ocena rzetelności sumarycznej skali osiągnięć matematycznych na poziomie rozszerzonym. Siłę efektów trzech czynników: efektu zróżnicowania jakości prac, egzaminatora oraz interakcji praca x egzaminator przedstawia wykres 5.



Wykres 5. Oszacowanie efektu zróżnicowania jakości prac, efektu egzaminatora oraz efektu interakcji praca x egzaminator, wskaźnik osiągnięć matematycznych, poziom rozszerzony, arkusz II

Jak widzimy, trzy słabsze „ogniwa" skali obniżyły rzetelność. Oczywiście 3% „szumu" to nie tragedia, ale — podobnie jednak jak w wypadku arkusza I — należy zdać sobie sprawę, że 97-procentowa rzetelność sytemu punktacji nie oznacza całkowitego braku rozbieżności w ocenie prac. Rozbieżności te są już w wielu wypadkach znaczne. Dla pracy rekordzistki nr 34 rozstęp między oceną maksymalną a minimalną wynosi 17 pkt. Jednak dla większości prac rozstęp ten jest mniejszy i w sumie średni rozstęp wynosi 6,9 pkt, co stanowi 12% teoretycznej zmienności wskaźnika sumarycznego (0–60 pkt). Również analiza odchyleń standardowych ocen pokazuje, że maksymalną wielkość efektu egzaminatora obserwujemy w wypadku pracy 34 — odchylenie standardowe ocen wynosi 5,9 pkt, przy średniej punktacji 32,4 pkt. Oznacza to przy

pierwszym sposobie kalkulowania 18% maksymalnej wartości, przy drugim — 28% maksymalnej wartości rozbieżności ocen. Z pewnością za część tych rozbieżności może odpowiadać niedoskonałość kopii, na których pracowali egzaminatorzy (choć zarówno w wypadku arkusza I, jak i II z analizy eliminowano zadania, w wypadku których błąd kopiowania mógł znacząco zaburzyć ocenianie), ale obniżenie rzetelności z pewnością wynika również z bardziej złożonej struktury zadań użytych w arkuszu (nie mylić z trudnością — zadanie może mieć prostą strukturę, a być trudne i *vice versa*) i związanych z tym problemów z konstrukcją modeli odpowiedzi.

Trzy punkty procentowe, których brakuje do doskonałości, z pewnością da się uzyskać poprzez sygnalizowane w poprzednim rozdziale procedury budowy modeli odpowiedzi, ale również warto przemyśleć, czy nie lepiej byłoby dla rzetelności — bez szkody dla trafności — arkusz II zbudować z większej liczby zadań o prostszej strukturze (nie znaczy — łatwiejszych).

Reasumując analizę rzetelności systemów punktacji dla egzaminu z matematyki można powiedzieć, że zgodnie z oczekiwaniami rzetelność ta była przynajmniej wystarczająca, a w wypadku arkusza I — dobra. Nie oznacza to, że nie ma co poprawiać. Szczególnie należy poprawić metodykę tworzenia modeli odpowiedzi, zwłaszcza dla bardziej złożonych strukturalnie zadań.

Innym zagadnieniem jest natomiast problem ekonomii testowania osiągnięć matematycznych za pomocą zadań otwartych. Oczywiście, w wypadku poziomu rozszerzonego pewna liczna zadań otwartych wydaje się niezbędna, ponieważ niektóre bardziej zaawansowane kompetencje matematyczne z pewnością można bardziej trafnie zbadać za pomocą tej właśnie formy zadań, choć niekoniecznie za pomocą zadań-molochów. Warto jednak — szczególnie w wypadku poziomu podstawowego — przemyśleć stosowanie wielozadaniowych wiązek zadań zamkniętych różnego typu, niekoniecznie WW, badających bardziej elementarne kompetencje matematyczne. Pamiętajmy, że zadania otwarte są bardzo pracochłonne: zajmują wiele czasu zarówno egzaminowym (czas egzaminu), jak i są czasochłonne — czyli kosztowne — w sprawdzaniu. Należy pamiętać, że analizując rzetelność i trafność dydaktyczną skali nie można dokonywać porównań w planie: jedno zadanie otwarte *vs* jedno zadanie zamknięte. Przeciętnie jedno zadanie otwarte w teście przekłada się na kilka, a nawet kilkanaście zadań zamkniętych. Pytając o wartość testu składającego się z zadań otwartych *vs* **testowi zbudowanemu** z zadań zamkniętych, należy porównywać np. test 10-zadaniowy (zadania otwarte) *vs* test 50–100-zadaniowy (zadania zamknięte). Przy takim porównaniu „otwartość" nie musi koniecznie przechylić wagę. Stosowanie ekonomicznych i jednocześnie dobrze obmyślanych zadań zamkniętych może być w wypadku wielu kompetencji matematycznych jak najbardziej właściwym rozwiązaniem. Z pewnością dla egzaminu na poziomie podstawowym cennym źródłem inspiracji mogą być zadania z amerykańskiego SAT-I.

11. PODSUMOWANIE ANALIZY RZETELNOŚCI SYSTEMÓW PUNKTACJI

Systemy punktacji wypracowań z języka polskiego stosowane w 2002 r. — zarówno dla poziomu podstawowego, jak i rozszerzonego — mają bardzo niską rzetelność. Oszacowania siły poszczególnych efektów pozwalają stwierdzić, że ocena punktowa pracy maturzysty w mniej więcej w tym samym stopniu zależy od jej cech, co od cech egzaminatora. Rzetelność punktacji arkusza systemów wyniosła 55%, a arkusza III — 49%. Stosowanie takich systemów oceniania oznaczałoby, że stronniczość oceniania przez szkolnego nauczyciela zastąpiliśmy nie mniejszym, „losowym szumem" związanym z efektami egzaminatora.

Za niską rzetelność systemów punktacji wypracowań odpowiadają przede wszystkim skale szacunkowe. Spośród nich wyjątkowo katastrofalną rzetelność ma skala szczególnych walorów pracy. Wynik ten nie dziwi. Sam charakter skali czynił z niej usankcjonowaną „oazę" subiektywności, ale nie zmienia to faktu, że użycie takiej skali nie da się pogodzić z dążeniem do obiektywizmu. Również skala jakości kompozycji i stylu nie może być używana w obecnej postaci. Rozczarowuje też skala rozwinięcia tematu. Jej kryterialny charakter mógł dawać nadzieję na zdecydowanie wyższą rzetelność. Jest ona wyższa niż skal szacunkowych, ale i tak zdecydowanie poniżej poziomu minimum.

Na tle systemów punktacji wypracowań, procedury punktacji testu rozumienia czytanych tekstów wypadają lepiej, choć i tu poziom rzetelności jest zdecydowanie zbyt niski. Efekty egzaminatora to w tym wypadku 20% zmienności ocen. Musimy pamiętać również o dwóch sprawach. Po pierwsze, swą wyższą rzetelność skala zawdzięcza głównie tym zadaniom skali, które odnoszą się do najniższych kategorii taksonomicznych. Zadania głębiej wnikające w rozumienie tekstu zdecydowanie trudniej poddawały się rzetelnemu punktowaniu. Po drugie, arkusz II obarczony jest błędem rzetelności dydaktycznej związanym ze skrajnie dużym błędem „próbkowania".

System punktacji testu historycznego (arkusz I) ma dostateczną, choć niedoskonałą rzetelność. Natomiast w wypadku arkusza II z historii notujemy wyniki porównywalne do rzetelności systemów punktacji wypracowań z języka polskiego, więc i w tym wypadku należałoby powiedzieć, że stosowanie takich systemów oceniania oznaczałoby, że stronniczość oceniania przez szkolnego nauczyciela zastąpiliśmy „ruletką" związaną z efektami egzaminatora.

Reasumując analizę rzetelności systemów punktacji dla egzaminu z matematyki można powiedzieć, że zgodnie z oczekiwaniami rzetelność ta była przynajmniej wystarczająca, a w wypadku arkusza I — dobra.

Podsumowując efekty symulacji zastosowania metod poprawy *post factum* rzetelności systemów punktowania dla arkuszy egzaminacyjnych z języka polskiego i arkusza II z historii, można powiedzieć, że przyniosły one znaczące efekty. Uzyskana rzetelność nie może jeszcze zadowolić, ale jeżeli weźmiemy pod uwagę korzyści, które możemy uzyskać rezygnując z jawnie subiektywistycznych skal, dopracowując pozostałe skale oraz rozwijając kulturę oceniania kryterialnego poprzez doskonalenie sytemu szkoleń i procedur kontrolnych w trakcie realnego sprawdzania prac, to możemy z dość dużym prawdopodobieństwem prognozować osiągnięcie 90–95% rzetelności.

To nie jest poziom, którym możemy się zadowolić, ale rzetelność, która może być zaakceptowana.

Otrzymane wyniki pozwalają na sformułowanie kilku wskazań. Zacznijmy od rekomendacji o charakterze ogólnym.

Po pierwsze, należy zmienić sposób wyrażania wyniku egzaminu maturalnego. Zarówno liczba punktów surowych, jak i wskaźnik procentowy nie pozwalają na wykonywanie koniecznych procedur korekcji i harmonizacji wyników. Należy przyjąć formułę skali punktów przeliczeniowych, np. od 0 do 100 pkt przeliczeniowych.

Po drugie, ponieważ problemy kontroli trudności skal punktacji oraz zapewnienia dostatecznej rzetelności systemów punktacji pytań otwartych nie dadzą się zadowalająco rozwiązać w fazie badań standaryzacyjnych, potrzebne są procedury *post factum* radzenia sobie z tymi wyzwaniami. Fakt ten wraz z koniecznością rozbudowania procedur kontroli procesu oceniania zdecydowanie wydłuża czas potrzebny na przeprowadzenie akcji maturalnej. Również potrzeba unikania przeciążenia osób zaangażowanych w egzaminowanie — które w oczywisty sposób przekłada się na obniżenie ogólnie rozumianej rzetelności oceniania — nakazuje zmianę harmonogramu. Można rozpatrywać dwa rozwiązania. Pierwsze to rozciągnięcie procedur oceniana wewnątrzszkolnego (tzw. egzaminów ustnych) na cały semestr letni klasy trzeciej/czwartej oraz przesunięcie egzaminów zewnętrznych (wg nazewnictwa *Rozporządzenia* tzw. pisemnych) na marzec/kwiecień. W tym wypadku należałoby zrezygnować z tradycji kończenia regularnych zajęć szkolnych przed maturą. Regularne, choć dopasowane w formie do potrzeb czasu maturalnego, zajęcia dla maturzystów kończyłby się w czerwcu. Zwróćmy uwagę, że tradycyjne kończenie zajęć przed maturą opiera się na niewysłowionym założeniu, że szkoła nie jest dobrym miejscem do przygotowywania się do ważnych egzaminów. Drugie możliwe rozwiązanie to przeniesienie akcji rekrutacji na studia wyższe na przełom sierpnia i września. Choć nie zyskamy tym pomysłem sympatii naszych kolegów — akademików, to sądzimy, że jest to propozycja warta przeanalizowania.

Przejdźmy do zaleceń szczegółowych.

Język polski i historia. Jak można poprawić jakość systemów punktowania w wypadku wypracowań z języka polskiego i arkusza II z historii? Oczywiście przede wszystkim przestrzegając zasad dobrej budowy skal i sprawdzając rzetelność systemów punktacji w fazie badań standaryzacyjnych. Ale w wypadku narzędzi egzaminacyjnych wykorzystujących złożone, słabo ustrukturowane zadania, koniecznym warunkiem uzyskania zadowalającej rzetelności punktowania jest stosowanie dodatkowo dwóch metod *post factum* podwyższania rzetelności oceniania. Pierwsza metoda — nastawiona na eliminację prostego efektu egzaminatora — polega na wprowadzeniu współczynników korekcyjnych; druga — nastawiona na redukcję efektu interakcji egzaminator x praca — polega na sprawdzaniu każdej pracy przez większą liczbę egzaminatorów (minimum dwóch). Bez wdrożenia tych metod nie sposób utrzymywać, że egzaminy zewnętrzne wykorzystujące złożone zadania otwarte są obiektywne i dostarczają porównywalnych ocen. W świetle rezultatów symulacji korzyści płynących z zastosowania metod *post factum* podwyższania rzetelności systemów punktacji, zastosowanie metody korekcji ocen egzaminatorów oraz zdublowanie liczby egzaminatorów jest racjonalne, a wzrost kosztów ma szanse przełożyć się na znaczący postęp w zakresie

obiektywności oceniania wypracowań maturalnych z języka polskiego, arkusza II z historii oraz — przez analogię — w wypadku wszystkich narzędzi egzaminacyjnych wykorzystujących złożone, słabo ustrukturowane zadania otwarte.

W wypadku testu czytania rozumianych tekstów najważniejsze jest natomiast zwiększenie liczby tekstów, zastąpienie części pytań krótkiej odpowiedzi zadaniami zamkniętymi oraz dopracowanie w fazie badań standaryzacyjnych modeli odpowiedzi dla zadań krótkiej odpowiedzi. Spełnienie tych zaleceń pozwoli zbudować satysfakcjonujący test rozumienia czytanych tekstów.

W wypadku testu historycznego (arkusz I) główną metodą poprawy rzetelności punktacji powinno być dopracowanie modeli odpowiedzi dla pytań krótkiej odpowiedzi i/lub ograniczanie liczby tego typu pytań do sytuacji, gdzie wymaga tego trafność.

Matematyka. Zadowalająca rzetelność systemów punktacji arkuszy matematycznych nie oznacza, że nie ma co poprawiać. Szczególnie należy poprawić metodykę tworzenia modeli odpowiedzi, zwłaszcza dla bardziej złożonych strukturalnie zadań. Innym zagadnieniem jest natomiast problem ekonomii testowania osiągnięć matematycznych za pomocą zadań otwartych. Oczywiście w wypadku poziomu rozszerzonego pewna liczna zadań otwartych wydaje się niezbędna, ponieważ niektóre bardziej zaawansowane kompetencje matematyczne z pewnością można trafnie zbadać za pomocą tej właśnie formy zadań, choć niekoniecznie za pomocą zadań-molochów. Warto jednak — szczególnie w wypadku poziomu podstawowego — przemyśleć stosowanie wielozadaniowych wiązek zadań zamkniętych różnego typu, niekoniecznie wielokrotnego wyboru, badających bardziej elementarne kompetencje matematyczne. Z pewnością dla egzaminu na poziomie podstawowym cennym źródłem inspiracji mogą być zadania z amerykańskiego SAT-I.

12. ANALIZA TRAFNOŚCI SYSTEMÓW PUNKTACJI WYPRACOWAŃ Z JĘZYKA POLSKIEGO WYKORZYSTYWANYCH NA MATURZE 2002

Czy stosując systemy punktacji, które — ze swej istoty — jakość wypracowania „rozbijają" na dziesiątki niezależnych kryteriów, można trafnie ocenić wartość wypracowania? Sceptycy mówią, że jest to niemożliwe, bo jakość tak złożonych wytworów to cecha synergiczna, dla której kluczowe są efekty całości, efekty kontekstowe, których nie może pomierzyć „atomistyczny" system punktacji. Czy ceną, jaką przyjdzie zapłacić za powtarzalność oceny, nie będzie zatem jej nietrafność? Na inny aspekt tego problemu zwraca uwagę jeden z naszych ekspertów:

Wprowadzenie klucza interpretacyjnego [tzn. schematu punktacji] *do matury z języka polskiego jest najgorszym z możliwych rozwiązań. Czemu najgorszym? Przede wszystkim dlatego, że konieczność istnienia klucza do sprawdzania wypracowań wpływa na jakość tematów. Klucz sprawia, że tematy muszą być odtwórcze, a nie problemowe. W efekcie nie dają one maturzyście szansy na polemizowanie, wyrażanie własnego zdania, popisanie się erudycją, bo tego przecież nie da się standaryzować w sposób, jaki proponują autorzy nowej matury.*

Czy opisywany przez eksperta problem jest przyczyną słabości tematów wypracowań, o której pisali nasi sędziowie kompetentni? Być może. W tym miejscu zajmie-

my się jednym tylko problemem: jak mają się oceny wystawiane przez egzaminatorów, „skrępowanych" atomistycznym schematem punktacji, do ocen osób mogących swobodnie, również zgodnie z zasadą synergii, kształtować swoje globalne oceny. Innymi słowy: czy zza tych drzew widać jeszcze las?

Oczywiście, jeżeli to przedsięwzięcie ma mieć sens, oceny globalne muszą być dokonywane przez osoby, których ocenom możemy zaufać. To istota metody, którą zastosujemy, tzn. metody sędziów kompetentnych.

Postanowiliśmy dobrać dwie grupy sędziów. Pierwsza miała reprezentować osąd polonistyki akademickiej. Znalazło się w niej 4 polonistów ze stopniem naukowym doktora. Niektórzy z nich mieli doświadczenie w szkolnym nauczaniu.

Druga grupa to doświadczeni, uznawani w swoich środowiskach za bardzo dobrych, nauczyciele języka polskiego. Przyjęliśmy zasadę, że nie mogli być oni wcześniej szkoleni na egzaminatorów maturalnych.

Prosiliśmy, by każdy sędzia kompetentny uporządkował prace zgodnie ze swoim rozumieniem kryteriów jakości szkolnych wypracowań. Przypominaliśmy tylko, że poziom podstawowy służył głównie podsumowaniu kształcenia polonistycznego, a poziom rozszerzony w większym stopniu miał być podstawą decyzji rekrutacyjnych przy przyjmowaniu na studia wyższe. Prosiliśmy w związku z tym o ewentualne — jeżeli sędziowie uznają to za adekwatne — zróżnicowanie perspektyw oceny. Prosiliśmy sędziów, by pracowali całkowicie niezależnie i nie porozumiewali się w trakcie sprawdzania prac.

Sprawdzenie wypracowań polegało na ich porządkowaniu. Prosiliśmy każdego sędziego o stworzenie dwóch oddzielnych rankingów: dla arkusza I i dla arkusza III. Jeżeli jakieś prace miały zdaniem sędziów taką samą jakość, tworzyli z nich grupkę prac o tej samej pozycji w rankingu.

By wyniki porządkowania prac poszczególnych sędziów były porównywalne, pozycje w rankingu przekształcano w tzw. rangi związane. Taki sposób wyrażania oceny sędziów kompetentnych wytrącał prosty efekt egzaminatora (por. poprzedni rozdział).

Ponieważ w dalszych analizach będziemy się posługiwać dwoma uśrednionymi ocenami wypracowań — nauczycielską i akademicką, przyjrzyjmy się, o ile oceny sędziów były ze sobą zgodne. Innymi słowy, zgodnie z metodologią opisaną w poprzednim rozdziale, przyjrzymy się rzetelności punktacji w obu grupach sędziów kompetentnych. W modelu nie uwzględniamy prostego efektu egzaminatora, ponieważ przy zastosowanej rangowej formule oceniania z założenia jest on zerowy.

Mimo porządkowego charakteru danych, by uzyskać oszacowania porównywalne z wynikami z analiz rzetelności punktowania, zastosowaliśmy model analizy wariancji. Może to budzić wątpliwości, ale zauważmy, że dane porządkowe w ten sposób uzyskane są i tak bardziej „metryczne" niż np. oceny na 4-pozycyjnej skali szacunkowej.

Tabela 6. Oszacowanie zgodności ocen wypracowań przez sędziów kompetentnych. Procent wariancji wyjaśnianej przez czynnik jakości pracy

Skala lub arkusz	Efekt zróżnicowania jakości prac w grupie sędziów-nauczycieli	Efekt zróżnicowania jakości prac w grupie sędziów-pracowników naukowych	Efekt zróżnicowania jakości prac w całej grupie sędziów kompetentnych
Arkusz I (wypracowanie, poziom podstawowy)	62%	47%	47%
Arkusz III (wypracowanie, poziom rozszerzony)	62%	48%	47%

Wyniki zgodności ocen w grupach sędziów kompetentnych są niestety bardzo kłopotliwe — szczególnie w grupie akademickiej. Zacznijmy od sędziów — nauczycieli. Zgodność ocen rzędu 62% nie jest duża. To poziom porównywalny z rzetelnością punktacji dla egzaminatorów (po wytrąceniu prostego efektu egzaminatora). Jest to wynik bardzo znamienny i pokazuje, że oceny globalne sędziów kompetentnych są niewiele mniej rzetelne niż dyscyplinowane skomplikowanymi schematami punktacji oceny egzaminatorów. Ale z drugiej strony możemy powiedzieć, że kto wierzył w wartość intuicyjnego, globalnego oceniania przez świetnych nauczycieli — fachowców, ten musi zrewidować swoją wiarę.

Jeszcze bardziej rozbieżne są oceny wystawione przez sędziów — akademików. Być może zabrakło w tym wypadku uniformizującego wpływu szkolnej kultury oceniania. Zgodność rzędu 47% właściwie stawia pod znakiem zapytania sens sumowania tych ocen i tworzenia jednego wskaźnika jakości prac prezentującego punkt widzenia polonistów — akademików. Co odpowiada za taki rozrzut ocen? Trudno wątpić w wiedzę polonistyczną naszych sędziów kompetentnych. To raczej dowód na to, że oceny globalne są silniej zakorzenione w subiektywnych, z definicji partykularnych, kryteriach oceny. Mimo wątpliwości, tworzymy wskaźnik sumaryczny. Ciekawe jest jednak, jak spojrzenie na jakość wypracowań akademików ma się do ocen nauczycieli — sędziów i egzaminatorów.

Zacznijmy od arkusza I. Poniższa tabela pokazuje, w jakim stopniu oceny poszczególnych egzaminatorów są skorelowane z naszymi kryteriami trafności — uśrednionymi ocenami sędziów kompetentnych.

Tabela 7. Korelacje ocen poszczególnych egzaminatorów z uśrednioną oceną dwóch grup sędziów kompetentnych (wartości współczynników korelacji Pearsona). Język polski, arkusz I

Numer egzaminatora	Sędziowie nauczycielscy	Sędziowie akademiccy
Egzaminator nr 1	0,71	0,56
Egzaminator nr 2	0,59	0,37
Egzaminator nr 3	0,53	0,46
Egzaminator nr 4	0,65	0,41
Egzaminator nr 5	0,68	0,60
Egzaminator nr 6	0,68	0,50
Egzaminator nr 7	0,70	0,64
Egzaminator nr 8	0,67	0,33
Średnia ocen egzaminatorów	0,80	0,60

Korelacja między średnią ocen sędziów — nauczycieli i sędziów — akademików wyniosła 0,71.

Zacznijmy od spostrzeżenia, że zgodnie z przewidywaniami oceny egzaminatorów są zdecydowanie silniej powiązane z opiniami sędziów — nauczycieli niż sędziów — akademików. W wypadku pierwszego kryterium możemy stwierdzić dość dużą, choć oczywiście daleką od doskonałości, zgodność. Korelacja rzędu 0,80 oznacza, że procent zmienności wspólnej tych dwóch ocen — egzaminatorów i sędziów kompetentnych — nauczycieli — wynosi 64%.

Znacznie niższą korelację obserwujemy w wypadku sędziów — akademików. Korelacja rzędu 0,60 oznacza, że oceny egzaminatorów tylko w 36% pokrywają się z tym kryterium. Dlaczego w wypadku drugiej grupy sędziów ta zgodność jest dramatycznie niższa? Naiwna odpowiedź brzmi — to dowód nietrafności systemów punktacji. Ale musimy pamiętać, że oceny akademików były bardzo rozstrzelone, innymi słowy wskaźnik oceny sędziów-akademików jest mniej rzetelny w psychometrycznym sensie. Jeżeli tak, to tak niska korelacja jest w jakieś części pochodną tej niskiej rzetelności.

Zanim przejdziemy do poziomu rozszerzonego, przyjrzyjmy się jeszcze wynikom dla poszczególnych egzaminatorów. Korelacje z ocenami sędziów są bardzo zróżnicowane, a szczególnie duże rozbieżności obserwujemy w wypadku korelacji z ocenami sędziów — akademików. Wartości współczynników w tym wypadku wahają się od 0,33 do 0,64, czyli od ok. 11% do 41% wariancji wspólnej ocen egzaminatorów i sędziów. To jeszcze jeden argument na rzecz dużej zmienności interpretacji systemu punktacji przez poszczególnych egzaminatorów.

Przyjrzyjmy się analogicznym wynikom dla poziomu rozszerzonego.

Tabela 8. Korelacje ocen poszczególnych egzaminatorów z uśrednioną oceną dwóch grup sędziów kompetentnych (wartości współczynników korelacji Pearsona). Język polski, arkusz III

Numer egzaminatora	Sędziowie nauczycielscy	Sędziowie akademiccy
Egzaminator nr 1	0,65	0,58
Egzaminator nr 2	0,54	0,40
Egzaminator nr 3	0,42	0,41
Egzaminator nr 4	0,79	0,75
Egzaminator nr 5	0,80	0,69
Egzaminator nr 6	0,67	0,79
Egzaminator nr 7	0,48	0,44
Egzaminator nr 8	0,56	0,62
Średnia ocen egzaminatorów	0,77	0,73

Korelacja między średnią ocen sędziów — nauczycieli i sędziów — akademików wyniosła 0,72.

Wyniki dla poziomu rozszerzonego są podobne jak w wypadku arkusza I. Jedyna różnica to mniejsza dysproporcja między korelacjami ocen egzaminatorów z sędziami — nauczycielami i sędziami —akademikami.

Czy w świetle otrzymanych wyników możemy jednoznacznie odpowiedzieć na pytanie o trafność systemu punktacji wypracowań z języka polskiego? Wydaje się, że nie można sformułować jednoznacznej diagnozy. Pierwsza konstatacja jest natury metodologicznej. Niska zgodność ocen wstawianych przez sędziów kompetentnych — szczególnie w wypadku pracowników naukowych wydziałów polonistycznych — każe wątpić w użyteczność tej metody. Po drugie, duże zróżnicowanie ocen sędziów kompetentnych każe porzucić wiarę w możliwość jednoznacznej oceny wypracowań dzięki li tylko wysokim kompetencjom merytorycznym oceniających. Szczególnie duże rozbieżności ocen grupy sędziów o najwyższych kompetencjach polonistycznych wskazują, że zależność może być wręcz odwrotna. Po trzecie, uśrednione oceny sędziów kompetentnych słabo lub umiarkowanie korelują z wynikami oceniania kryterialnego. To oznacza, że globalna ocena jest powiązana z oceną „zatomizowaną" — i to wniosek optymistyczny — ale związek ten jest słaby i bardzo zmienny od egzaminatora do egzaminatora. W wypadku ocen wystawianych przez egzaminatorów, dla których korelacje są rzędu 0,30, należałoby stwierdzić, że oceny te nie są trafne. Dla tych, dla których korelacje wynoszą 0,70 i więcej, można by bronić tezy o zadowalającej trafności diagnostycznej.